DOI:10.15923/j.cnki.cn22-1382/t.2018.1.14

面向互联网的隐私保护关键算法

杨秀菊

(泉州信息工程学院,福建 泉州 362000)

摘 要: 针对离散点消除问题,通过聚类裁剪的离散点检测算法有效提高检测效率。针对个性化匿名问题,引入个性化属性权重密度聚类匿名算法以消除离散点影响。

关键词: 数据隐私保护;个性化;聚类;k-匿名

中图分类号: TM 417 **文献标志码:** A **文章编号:** 1674-1374(2018)01-0080-05

An internet privacy protection key algorithms

YANG Xiuju

(Quanzhou Institute of Information Engineering, Quanzhou 362000, China)

Abstract: To eliminate discrete points, a discrete point detection algorithm based on clustering clipping is applied to improve the detection efficiency. To solve the problem of personalized anonymity, individual attribute weight density clustering anonymous algorithm is introduced to eliminate the influence of discrete points.

Key words: data privacy protection; personalized; clustering; k-anonymity.

0 引 言

针对传统数据匿名隐私保护技术中存在信息损失过大[1]、数据效用低下等问题,越来越多的数据挖掘技术被引用进来[2]。基于频繁项集的关联规则挖掘和聚类挖掘就是很好的应用范例。由于不同的聚类方法在处理不同类型、不同规模的数据各有其不同的优缺点,而基于聚类的匿名隐私技术实际上受聚类本身的制约也比较大[3],如何根据实际选择比较符合的聚类方法、尽量避免因为聚类本身缺陷带来的相应问题便成为了研究的重点之一。

1 基于 k-PROTOTYPE 裁剪的离散点检测算法

离散点又称离群点、异常点,对数据挖掘等数据统计处理技术产生显著影响,同样,在基于聚类的匿名算法中,如果未能很好地处理离散点,误将其分入某个簇中,就有可能导致该簇的过度泛化,从而使信息损失变大。

常用的离散点检测方法有:基于统计的检测、

收稿日期: 2017-11-17

基金项目: 福建省中青年教师教育科研项目(JAT160612)

作者简介: 杨秀菊(1971—),女,汉族,黑龙江哈尔滨人,泉州信息工程学院讲师,主要从事软件工程方向研究,E-mail:84194459@qq.com.

基于邻近度的检测、基于密度的检测。

基于密度的离散点检测算法基本思想是：不将离散点看做一种简单的二元对象，而是用一个权值来评估它的离散度。这个离散度叫局部离散因子(LOF)，表示该对象相对于其附近领域的孤立情况。

通过计算离散度来检测离散点的方法叫LOF算法，这也是基于密度的离散点检测的代表算法。这种方法可以同时检测出全局离散点和局部离散点。

1.1　基于密度的离散点检测算法的初步改进

LOF算法在查询某对象 A 的第 k 距离领域时，实际上存在效率问题，因为在对 A 查询完领域以后，这些信息都会被放弃，要查询其领域内其他对象的领域，又必须重新开始对该对象计算第 k 距离和第 k 距离领域。但事实上，这些对象的领域与 A 的领域有很大的重叠。

领域递推示意图如图1所示。

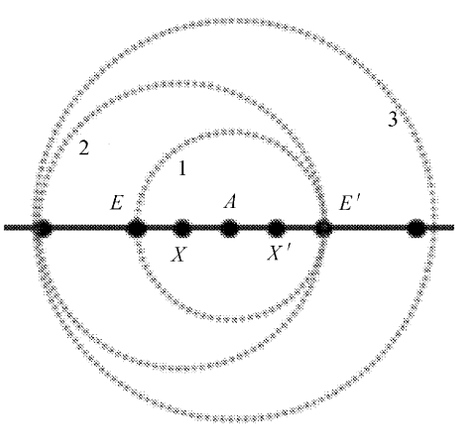

图1　领域递推示意图

区域1与区域2内切，区域2与区域3内切。设 X 为 A 的领域中的某点，而 E 为 A 的领域边界上的点，则以 A 为圆心、k-distance(A)为半径的区域1包含了 A 领域中的所有点并且除去 A 本身的个数为 k。以此类推，则对 A 领域中某点 X，其领域中的所有点一定包含在以 A 为圆心，$2*d(X,A)+k-\text{distance}(A)$ 为半径的区域中，这便省去了对 X 的许多“无关点”的考察。

1.2　基于 k-PROTOTYPE 裁剪的 LOF 算法

使用聚类进行数据预处理，实现非离散点的预先剪裁，从而有效降低LOF算法的样本集合。为了提高整个算法的效率，预先剪裁只是为了剔除高内聚的点，对聚类质量要求并不高，因此，使用划分聚类中的经典算法 k-均值算法的扩展算法 k-PROTOTYPE 来处理混合属性数据(k-均值算法只能处理数值属性)。

1.2.1　k-PROTOTYPE 算法的基本定义

在改进算法 k-PROTOTYPES 中定义了一个对数值与分类属性都适用的相似度的度量方法，根据此方法对数据集进行聚类，以获得最优聚类结果。

下面给出 k-PROTOTYPES 算法的相关定义：

定义 1　相异度。设数据集 U 同时包含对数值与分类属性，X、Y 为 U 中的两个对象，其中

$$X=(x_1,x_2,\cdots,x_p,x_{p+1},\cdots,x_m)$$

$$Y=(y_1,y_2,\cdots,y_p,y_{p+1},\cdots,y_m)$$

$$m>p$$

且前 p 个为数值属性，$p+1$ 到 m 为分类属性，则对象 X、Y 之间的相异度公式为：

$$d(X,Y)=\sum_{i=1}^{p}(x_i-y_i)^2+\lambda\sum_{i=p+1}^{m}\delta(x_i,y_i)$$

其中

$$\delta(x_i,y_i)=\begin{cases}0, & x_i=y_i\\ 1, & x_i\neq y_i\end{cases}$$

定义 2　成本函数

$$P(X,Y)=\sum_{j=1}^{k}\sum_{i=1}^{p}(x_i-y_j)^2+\lambda\sum_{j=1}^{k}\sum_{i=1}^{p}\delta(x_i,y_i)$$

1.2.2　改进初始聚类中心选取的 k-PROTOTYPE 算法

在使用 k-PROTOTYPE 算法剪裁 LOF 之前，需要对初始聚类中心的选取做一些改进，利用等分区间思想，具体步骤如下：

1)设定数据集中心点，数值属性取平均值，分类属性取众数；

2)找到一个离中心点最远的点，记为第一个初始点 A_1，最远距离设为 r；

3)在剩余点中找到与 A_1 距离最接近 $r/(k-1)$ 的点，记为第二个初始点 A_2；

4)以此类推，在剩余点中找到与 A_1 距离最接近 $(i-1)r/(k-1)$ 的点，记为第 i 个初始点 A_i，直到 $i=k$，全部 k 个初始点都找到。

由此可得改进初始聚类中心选取的 k-PROTOTYPE 算法的详细伪代码。

2 密度聚类的个性化属性权重匿名算法[4]

2.1 基于OPTICS的个性化属性权重匿名算法

2.1.1 算法的基本步骤

目前能实现隐私保护[5]的隐私模型主要是以k-匿名模型为原型进行的拓展与优化[6]。而在实现个性化属性权重的匿名算法时，同样选取k-匿名模型作为原型，以泛化作为主要的匿名方法。基于OPTICS的个性化属性权重匿名算法基本步骤如下：

1)数据集预处理，生成准标识符数据表与每个分类属性的层次分类树；

2)进行离散点检测，生成处理过的数据表与离散点集；

3)确定个性化属性权重，从而得出加权距离计算公式，确定信息损失度量标准；

4)使用OPTICS进行聚类，计算生成聚类的信息损失，然后进行聚类调整；

5)泛化每个簇中的所有准标识符属性值。

2.1.2 个性化属性权重距离公式

样本相似度或距离度量是进行聚类前必须预先确定的。由于本算法支持对混合型数据集的处理，因此，需要分数值属性和分类属性进行距离定义，最后确定个性化属性权重距离的计算公式。

2.1.2.1 数值属性的距离

设N是一个数值属性，两个元组对应该属性的属性值分别为x和y，则它们的距离为：

$$dN(x,y)=\frac{|x-y|}{N_{\max}-N_{\min}}$$

2.1.2.2 分类属性的距离

设D是一个分类属性，两个元组对应该属性的属性值分别为x和y，则它们的距离为：

$$dD(x,y)=\frac{\sum_{i=1}^{f_{xy}-1}W(i,i+1)}{\sum_{j=1}^{h-1}W(j,j+1)}$$

其中，$W(i,i+1)(1\leqslant i\leqslant h-1)$为层数$i$到层数$i+1$的权重，通常等于$i$。

2.1.2.3 二元属性的距离

设B是一个二元属性，两个元组对应该属性的属性值分别为x和y，则它们的距离为：

$$dB(x,y)=\begin{cases}0, & x=y\\ 1, & x\neq y\end{cases}$$

综合3种属性的距离计算方法，最后得到个性化属性权重距离公式。设两个元组X和Y，其中，前a个为数值属性，$a+1$到b为分类属性，$b+1$到c为二元属性，有

$$X=(x_1,x_2,\cdots,x_a,x_{a+1},\cdots,x_b,x_{b+1},x_{b+2},\cdots,x_c)$$

$$Y=(y_1,y_2,\cdots,y_a,y_{a+1},\cdots,y_b,y_{b+1},y_{b+2},\cdots,y_c)$$

则得到的个性化属性权重距离公式为：

$$d(X,Y)=\sum_{i=1}^{a}w_i*dN(x_i,y_i)+\sum_{i=a+1}^{b}w_i*dD(x_i,y_i)+\sum_{i=b+1}^{c}w_i*dB(x_i,y_i)\quad(1)$$

2.1.3 信息损失度量

隐私匿名算法中[3]常用的信息损失度量标准有基于泛化层次的信息损失度量方法、基于元组辨别度的信息损失度量方法和基于熵的信息损失度量方法等。在本算法中，使用抑制单元来衡量匿名代价。其定义如下：

$$C(T')=\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}HM(X_{ij},X'_j)}{n*m}$$

式中：$HM(X_{ij},X'_j)$——从X_{ij}泛化为X'_j，前后记录的海明(Hamming Distance)距离，也就是码距。

2.1.4 初始邻域半径参数ε的选取

前面说过，OPTICS是针对基本密度聚类算法DBSCAN的输入参数敏感性而进行的优化，然而仍旧需要输入初始ε，但不会对聚类结果产生太大影响。

3 实验与验证

3.1 实验数据与环境

实验选用UCI机器学习数据库中Adult数据集，Adult数据集包括两组数据集：训练集(Train)和测试集(Test)。根据实验结果，选择隐私保护效果比较好的QI10={Age,Capital Gain,Country,Education Num,Hours-per-week,Marital Status,Race,WorkClass}作为准标识符属性，其中属性{Age,Capital Gain,Education Num,Hours-per-week}为数值属性，其余为分类属性。实验环境为Intel(R)Core(TM)Duoi7-3632QM处理器、2.20 GHz主频，8 GB的内存，MicrosoftWindows7SP164位操作系统。编程语言为C++。

3.2　实验结果

3.2.1　基于 k-PROTOTYPE 裁剪的 LOF 算法

为了验证离散点检测算法的有效性，在实验中，分别对属性{Age，Capital Gain，Education Num，Hours-per-week，Race}中各添加了两个总共 10 个的扰动数据。其中 Age，Capital Gain，Education Num，Hours-per-week 为数值属性，Race 为分类属性。然后，从两个方面分别与原始 LOF 算法进行对比。设定预期离散点个数 $n=20$，数值属性与分类属性的相异度参数 $r=1$。

3.2.1.1　检测率与剪枝率

调整 k 值，重复 5 次实验。初始值设为 100，增长步长为 50。检测率与剪枝率见表 1。

表 1　k 变化时 OKPLOF 与 LOF 的检测率与剪枝率对比

k	$K'(*)$	检测率/%		剪枝率/%
		OKPLOF	LOF	
100	301	0.9	0.6	60.18
150	201	0.9	0.7	59.52
200	150	1.0	0.8	59.52
250	120	1.0	0.8	57.69
300	100	1.0	0.8	55.78

在本实验中，$k'=30\ 162/k$。

3.2.1.2　运行时间

5 次实验的运行时间对比见表 2。

表 2　k 变化时 OKPLOF 与 LOF 的运行时间对比

k	$K'(*)$	检测率/%	
		OKPLOF	LOF
100	301	30	93
150	201	41	102
200	150	51	116
250	120	59	135
300	100	65	157

3.2.2　基于 OPTICS 的个性化属性权重匿名算法

本实验使用抑制长度和运行之间来评估算法，并与基于密度的聚类匿名算法 DSAED 进行比较。为了统一标准，在与 DSAED 进行对比实验的过程中各准标识符的权值均设为 1。

3.2.2.1　个性化

主要通过 OBPA 算法的内部对比不同属性重要程度下的抑制单元数，见表 3。

设

$W=(0.132\ 9,0.080\ 5,0.265\ 8,0.520\ 8,0.258\ 3,0.637\ 0,0.212\ 4,0.424\ 6,0.104\ 7)$

表 3　不同属性重要程度下的抑制单元数对比

QI	抑制单元数	
	1(=0.132 9)	2(=0.265 8)
Age	10 658	6 148
Race	21 476	17 354

3.2.2.2　抑制单元

调整 k 值，重复 5 次实验。抑制单元数对比见表 4。

表 4　k 变化时的抑制单元数对比

k	抑制单元数	
	OBPA	DSAED
16	75 683	78 646
32	87 771	90 926
64	98 364	103 678
128	105 002	112 284
256	110 080	118 134

3.2.2.3　运行时间

5 次实验的运行时间对比与表 4 相似，表略。

3.3　结果评估

1)在第一组实验中，OKPLOF 的时间优化效果明显，同时从检测率上可以看出，OKPLOF 也解决了 LOF 算法不能处理混合属性的问题。

2)在第二组实验中,OBPA与DSAED均采用密度聚类,但OBPA的信息损失明显低于DSAED,而运行时间在k值大于32时也开始低于DSAED,优化效果得以证实。而在个性化实验中,代表属性重要程度的权值增加,对应准标识符的抑制单元数下降,说明权值使得等价类中该属性的属性值相似度更高,有效减少了泛化所带来的信息损失,个性化算法有效。

4 结 语

实现了基于密度聚类并消除离散点影响的个性化属性权重匿名算法,首先介绍了聚类和相关算法,然后针对离散点消除,对现成的LOF算法进行初步改进,并由此提出了通过基于优化初始点选取的k-PROTOTYPE聚类裁剪的离散点检测LOF算法,介绍了基于密度聚类的OPTICS算法,并阐述了属性权重的引入,确定了个性化属性权重距离公式,最后,提出了消除离散点影响的个性化属性权重密度聚类匿名算法,并用实验进行算法的对比验证,结果证明两个算法均合理有效。

参考文献:

[1] 王佳慧,刘川意,方滨兴.面向物联网搜索的数据隐私保护研究综述[J].通信学报,2016,37(9):142-153.

[2] 俞志斌,周彦晖.基于关键字的云加密数据隐私保护检索[J].计算机科学,2015,42(s1):132-136.

[3] 张晓琳,王萍,郭彦磊.社会网络子集个性化隐私保护策略[J].计算机应用研究,2015,32(10):3026-3029.

[4] 魏姣妲,逄焕利.基于区域中心点的多层次数据集密度聚类算法[J].长春工业大学学报,2016,37(6):576-580.

[5] 张小波,付达杰.网络信息资源个性化推荐中隐私保护的研究[J].软件,2015,36(4):62-66.

[6] 王良,王伟平,孟丹.FVS k-匿名:一种基于k-匿名的隐私保护方法[J].高技术通讯,2015,25(3):228-238.